Capstone Project - The Battle of Neighborhoods (Week 2)

1. Introduction

Assume an entrepreneur plans to open a fitness club in Los Angeles (LA) and she/he wants to know where she/he should choose to be the location of her/his club. The goal of this project is to solve this location problem and give him or her an idea about which area in LA would be suitable

There are totally 272 neighborhoods in LA. To simplify the analysis, we will reduce the number of neighborhoods to explore. Apparently, we want to have more people to become a member of the club and so firstly we want to consider the population factor of the neighborhoods. Therefore, we will only select the top populous neighborhoods to discuss.

2. Description of data

Data from three sources will be used:

1) All 272 neighborhoods with location coordinates, download from: https://usc.data.socrata.com/dataset/Los-Angeles-Neighborhood-Map/r8qd-yxsr

	Neighborhood	Longitude	Latitude
0	Acton	-118.169810	34.497355
1	Adams-Normandie	-118.300208	34.031461
2	Agoura Hills	-118.759885	34.146736
3	Agua Dulce	-118.317104	34.504927
4	Alhambra	-118.136512	34.085539

2) the population density of neighborhoods data in LA, retrieved from：
http://maps.latimes.com/neighborhoods/population/density/neighborhood/list/

	Neighborhood	Population Per SQMI	Total Population
0	Koreatown	42611	115070
1	Westlake	38214	103839
2	East Hollywood	31095	73967
3	Pico-Union	25352	42324
4	Maywood	23638	28083

3) venues data of these top neighborhoods explored from Foursquare API.

The first two data sets were combined to get coordinate information of the top neighborhoods that a population density higher than 10k for the later-on venue data extraction from Foursquare. There was not much data cleaning needed other than selection of interested neighborhoods that has population density higher than the 10k threshold in dataset 2. However, the venue data extracted from Foursquare were further processed by converting the category variables into numeric labels for applying clustering algorithm. Then, the neighborhoods will be clustered according to their similarity in venue categories using K-means algorithm. Lastly, the clustering results will be visualized after variable reduction using principle component analysis. After this analysis, we should be able to select some neighborhoods that are more suitable than others to start up the new fitness club.

3. Method and Results

3.1 Clustering

I accessed to venue information in each of the 95 neighborhoods through Foursquare API, which including information of venue categories and coordinates. There were 249 unique venues in all the 95 neighborhoods. In order run clustering analysis, the data was first grouped and sum the total number of each venue in each neighborhood. The proportion of each venue was also calculated and then was ranked in each neighborhood, then I got top ten venues in each neighborhood. The calculated proportion data was used for k-mean clustering.

There was no clear elbow point in the plot (Figure 1), I chose the k=5 as algorithm parameter. In order to visualize the clustering results, I applied Principal Component Analysis (PCA) to reduce the number of features (variables), as shown in Figure 2. Apparently, the first two components could sufficiently explain the majority of the variance in our data. The 2D scatter plot with two principal components (Figure 3a) showed some overlap of points. For more clear visualization, I also generated a 3D using three principal components (Figure 3b), which gave better overview of the clustering results.

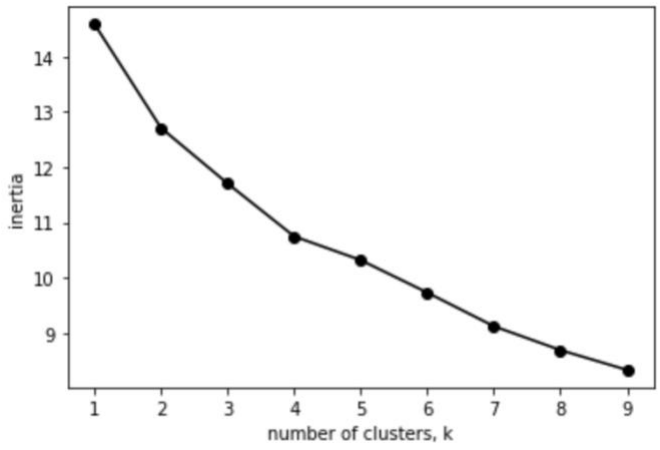

Figure 1. The elbow plot for determination of k.

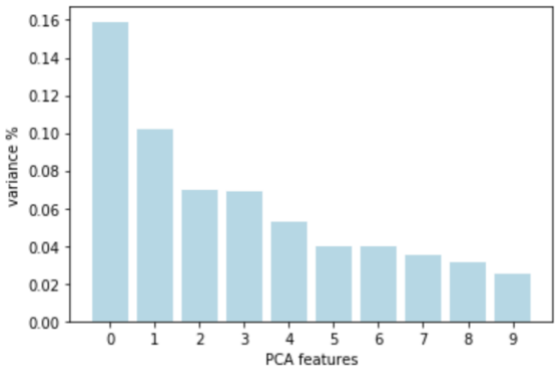

Figure 2. PCA results.

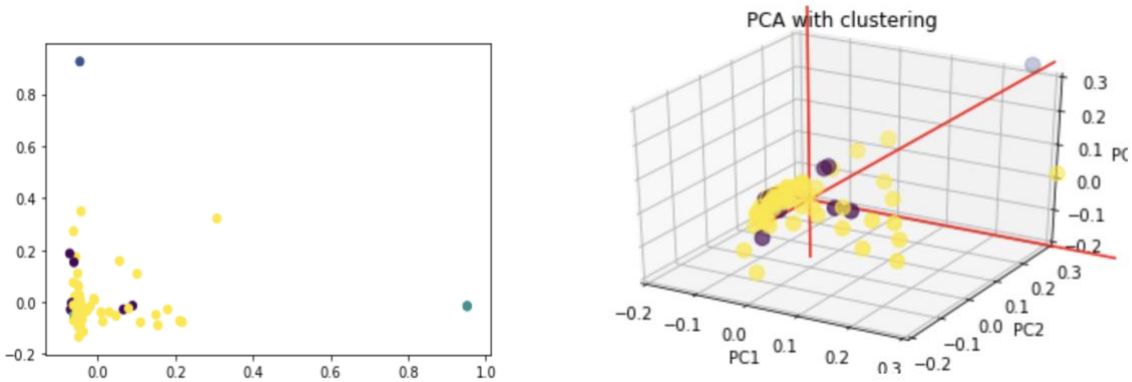

Figure 3. Scatter plots with principal components: (a) two PCs, (b) three PCs.

Each of clusters was further analyzed. The cluster 2, cluster 3 and cluster 4 have only one type of venues, namely, Burger Joint, Park, Speakeasy, respectively. So, it is easy to define these three clusters according to their venue categories. For cluster 1 and cluster 5, the top ten venues for each cluster were shown in the bar chart below (Figure 4). For cluster 1, the dominant venue is grocery store and other shops, while the cluster 5 mainly have restaurant and sorts of food stores.

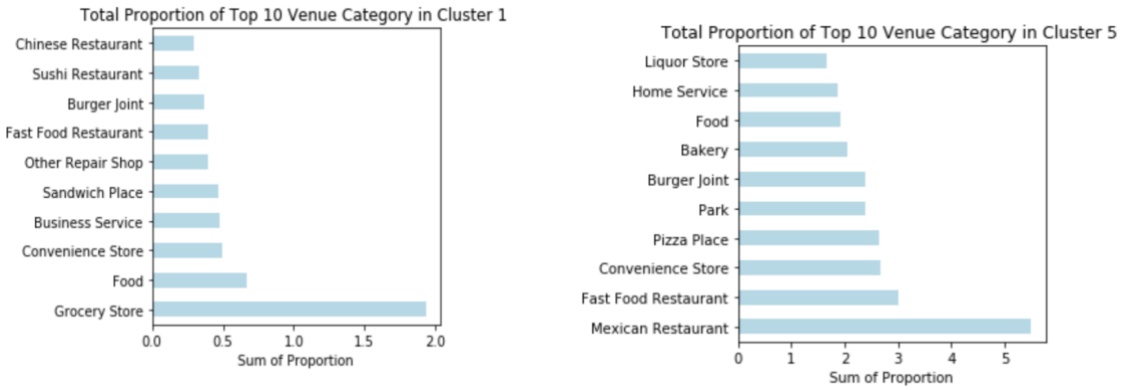

Figure 4. The top ten venues for each cluster.

To explore the venues that relevant to gym, fitness or yoga, I extracted all the venues that contain these three key words and found that only these fitness related venues are only in Cluster 5. Therefore, we should avoid those neighborhoods in cluster 5 to set up the new gym/fitness club.

	Venue	Total
32	Gym	0.625155
41	Gym / Fitness Center	0.510673
67	Yoga Studio	0.268912
160	Gymnastics Gym	0.058824

3.2 Map Visualization

After clustering, I generated a Folium map of LA with labeled clustered neighborhoods, we can see that the dominant cluster is Cluster 5 (Figure 5), which was also confirmed by individual cluster analysis above.

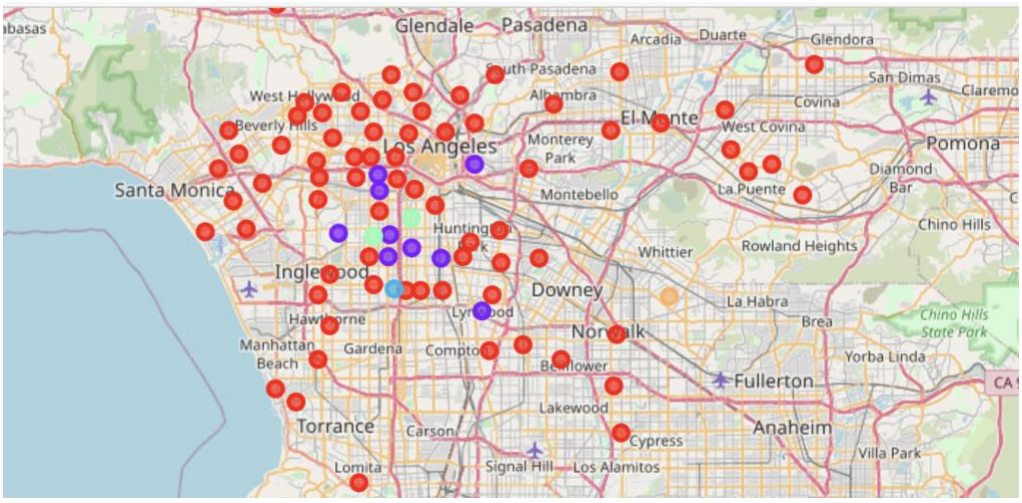

Figure 5. Folium map of LA with labeled and clustered neighborhoods (Purple points: Cluster 1; Blue point: Cluster 2; Green points: Cluster 3; Orange point: Cluster 4; Red points: Cluster 5)

By making an assumption that in area with higher population density, we could possibly have more members joining in the club if blocking other features, such as resident features and local economical features, etc. Then, I created a Choropleth map (Figure 6) based on the population density of the 95 neighborhoods and label the neighborhoods in selected clusters, i.e., cluster 1 to cluster 4, so that we can make further decision on neighborhoods selection according to population density. The potential candidates could be those neighborhoods in Cluster 1 (Purple points in Figure 5 and Figure 6).

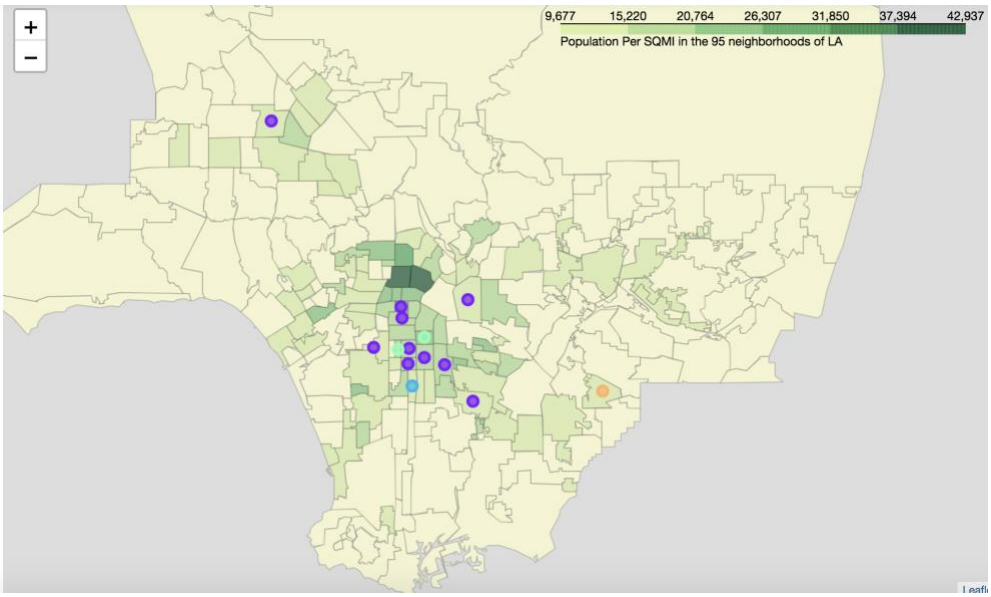

Figure 6 Choropleth map with labeled clusters.

4. Discussion and Conclusion

Clustering using K-means algorithm is a commonly approach and it is fairly fast. However, the clustering results might change each time when we run the algorithm since the initial k centroids are randomly chosen at beginning. Moreover, in our case that clustering based on venue categories and counts, it might be more reasonable to consider the similarity, other than the distance between means. This might be the reason why the 2D and 3D plot of the clustering results did not clear show distinction of each cluster. Therefore, future analysis might need to include other clustering approach that based on similarity of each data, such as spectral clustering algorithm.

In summary, from the analysis and visualization above, we can choose neighborhoods in cluster1 (purple dots in map), such as Adams-Normandie and Florence-Firestone to start the gym/fitness club business. From the Foursquare venues data and clustering analysis, we know there are no gym/fitness venues at all. Additionally, among all the neighborhoods that have no gym/fitness venues, they have relatively higher population density which might be beneficial for having more members. However, further analyses are also needed to obtain the exact location. To achieve this, we will still need to know the number of similar venues nearby the area, since the nearby neighborhoods which are located in cluster 5 might have numbers of fitness centers and we do not expect to choose a location that has large numbers of same venue nearby.